Intel® Xeon® Processor D: The First Xeon Processor Optimized for Dense Solutions

Dheemanth Nagaraj, Chris Gianos

Server Architecture

Acknowledgements:

Xeon-D Team

Legal Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804.

No computer system can be absolutely secure.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Available on select Intel® processors. Requires an Intel® HT Technology-enabled system. Your performance varies depending on the specific hardware and software you use. Learn more by visiting http://www.intel.com/info/hyperthreading.

Agenda

- Introducing Xeon® Processor D aka Broadwell-DE
- Architecture Overview and Features
  - Broadwell Core
  - Integrated Ethernet Features
  - Virtualization, RAS, Security, Power Management
  - Key Storage Features
- Performance Benchmarks
- Conclusions

Introducing Xeon® Processor D

- New Low Power product family optimized for workloads that scale with single socket nodes
- Brings Xeon Class RAS, Virtualization capabilities to the 20 – 45w design points
- Design focused on breakthrough Perf/Watt and Dense Form Factor Optimizations
- Integrates critical Server, Networking IOs and features for Communications and Storage usages

Where Xeon D Fits

Form Factors challenged on Board real-estate, Thermally constrained & requiring high Compute Density

Mid-Range Comms

Ex: Branch Office Routers, Security Appliances, Wireless RNC

Entry Storage

Ex: Cloud Warm Data/SMB Mini-ITX, Unified Storage (SAN and NAS)

Dense Hyperscale Cloud

EX: Dedicated hosting, Web Tier

Enhancements to Prior offerings

Client based low power CPUs, Xeon E3, Atom S1200 SoC

SoC Integration

Key IOs, Networking/Storage features

Performance & Scalability

Core Count, LLC size, Mem Capacity

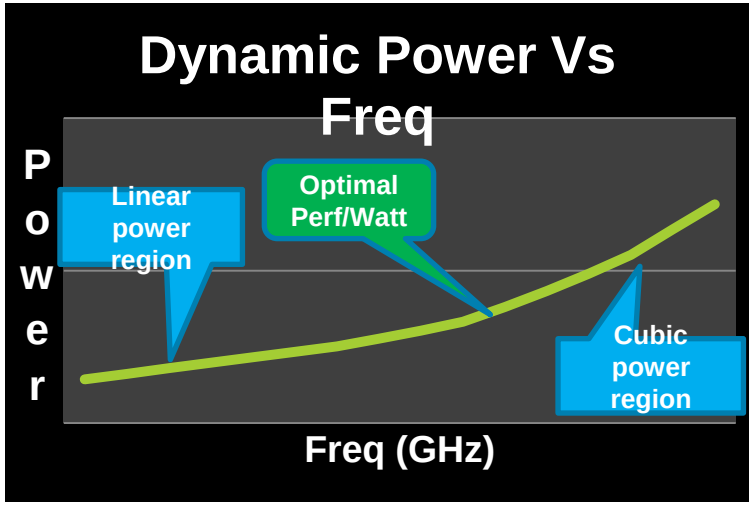
Single platform design across multiple price/perf points

Server class RAS & Virtualization

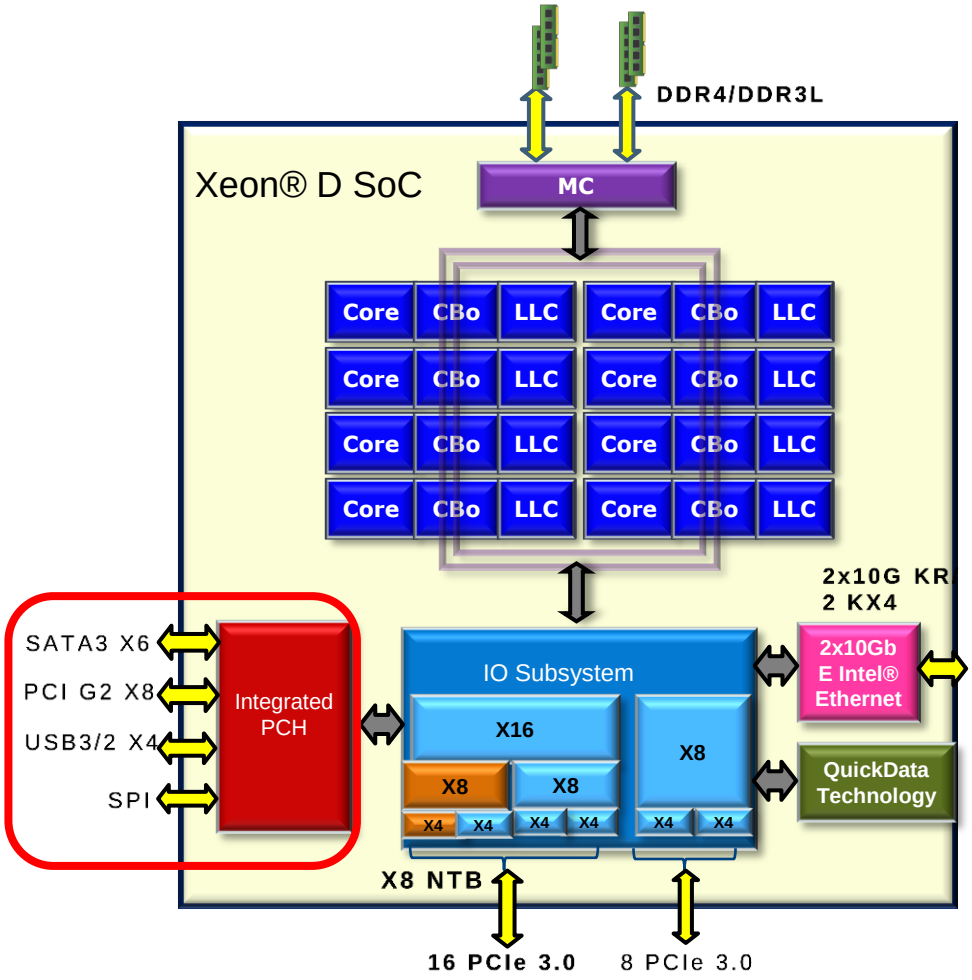
Memory/IO RAS, Full APIC-V, SRIOV

Design Optimizations

- Optimizations for node, overall rack Perf/Watt
  - Balanced Node => Compute vs. Memory Bandwidth vs. IO bandwidth
  - Operating point close to the knee of the Power vs. Freq curve => Optimal Perf/Watt
  - Leakage power significant piece of the pie at this operating point => Process variant optimized for low leakage
- Focus on compute Density
  - SoC integration of platform components
  - Choice of BGA package
  - TDP and platform thermal co-optimization
  - Reference design with 16 Nodes in 3U module; Up to 6 Nodes per 1U also possible



Intel® Xeon® Processor D – Block Diagram



- First Server SoC on Intel 14nm process
- TDP Range : 20 – 45W
- 8 Broadwell cores, 16 threads
- 8 slice shared Last Level Cache (L3)
  - 12MB total LLC
- 2 DDR4/DDR3L memory channels
- Integrated Ethernet with 2 X 10G KR/KX4 ports
- 24 PCIe Gen3, 8 PCIe Gen2 lanes
- 6 SATA Gen3, 4 USB3/2
- Integrated Boot, Legacy IO, Manageability Engine
  - SPI, SMBus, UART, LPC, GPIO, 8259, I/O APIC, 8254 Timer, RTC

LLC : Last Level Cache

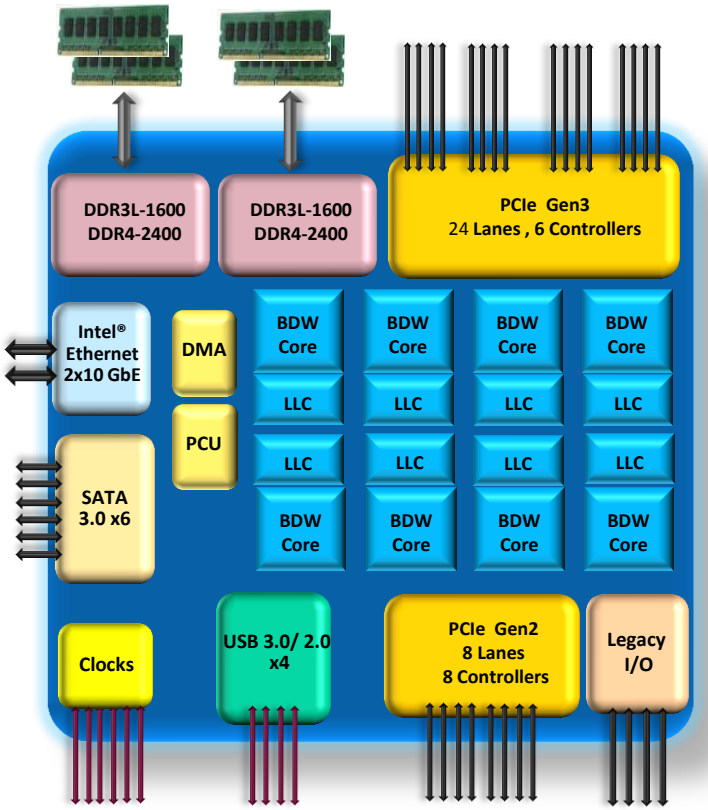
MC : Memory Controller

Cbo : Cache controller

NTB : Non Transparent Bridge

Baseline Architecture and Features

- CPU Cores
  - L1 cache: 32K Data/ 32K Instruction; L2 cache: 256K per core
  - Addressing: 46b Physical; 48b Virtual
- On-die interconnect & Last Level Cache
  - Bi-directional High BW ring interconnect
  - 12 MB Distributed Shared cache (1.5M/slice)
  - Latency: 21ns; Bandwidth: ~250GB/s
- Memory Speeds and Feeds
  - Speeds: DDR4 2400 MT/s; DDR3L 1600 MT/s
  - Latency: 66 ns page hit; 80 ns closed page
  - BW: 100% R 36.2 GB/s
    2R/1W 32.8 GB/s



Features Continued

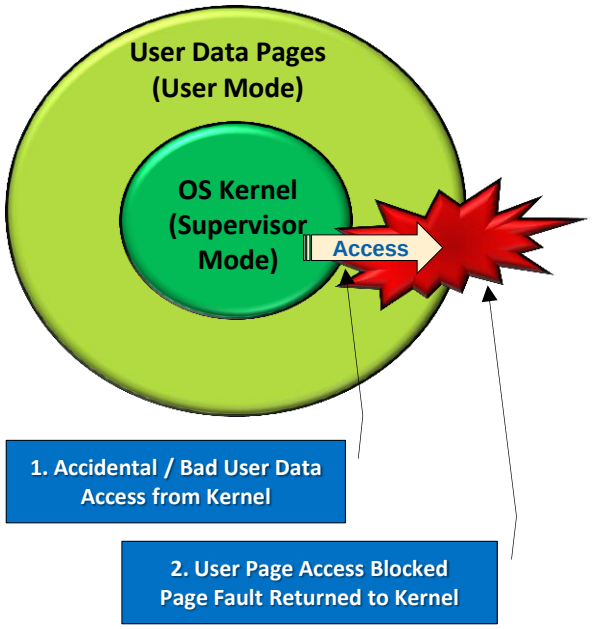
- Memory Capacity and RAS
  - RDIMMs up to 128GB; UDIMM/SoDIMM up to 64GB
  - Enhanced ECC w/ SDDC support; DDR4 CAP; Patrol/Demand Scrub; Data scrambling
- PCIe Subsystem and RAS
  - x24 PCIe Gen3 (6 controllers); x8 PCIe Gen2 (8 controllers)
  - eCRC (covers switches and bridges), Advanced Error Reporting; PCIe Hotplug
- Technologies
  - Intel® VT (VT-x, VT-D2), TXT, PECI over SMBUS, PSE
- Power Management
  - Per Core P-States (PCPS), Uncore Freq Scaling (UFS), Core RAPL
  - Hardware PM (HWPM)

Intel® Xeon® Broadwell Core

Evolution of Xeon® Haswell Core Architecture on 14nm

- Improved floating point performance
  - Radix-1024 divider: Decreased latency, increased throughput
  - Split scalar divider: Pseudo-double BW for scalar divider uops
  - Vector FP multiply latency decrease (to 3 cycles from 5)
- STLB and Page Miss Handling Improvements
  - Native 16- entry 1G STLB array; Increased 1.5 KB STLB
  - 2 simultaneous Page Walks enabled
- Other ISA improvements
  - ADC, CMOV, PCLMULQDQ, VCVTPS2PH
- Security Enhancements
  - Supervisory Mode Access Protection => See diagram
  - Faster ADC/SBB, ADCX/ADOX instructions => 30% improvement on RSA public key performance
  - RDSEED: Provides High quality seed values for Software pseudo-random number generators

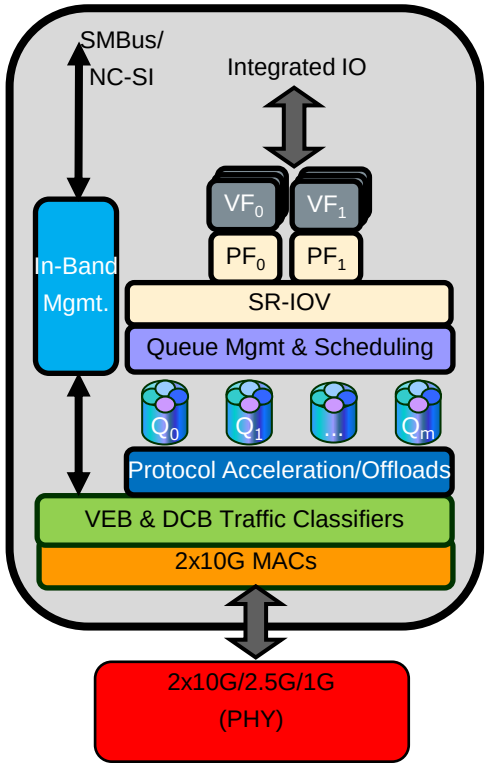
SMAP



Prevents unintended supervisory mode accesses to data on user page

Integrated 10 GbE Intel® Ethernet

- Dual Port 10GBe MAC supporting 1G/2.5G/10G
  - Support for Windows, Linux with single driver across SoCs, Chipsets and discrete NICs
- Standards based Virtualization support
  - SRIOV (64 Virtual Functions), VMDq (64 VMs), 128 Tx/Rx queues per port, Virtual Ethernet Bridge
- High performance Unified Networking support
  - Data Center Bridging supporting 8 traffic classes for prioritized flow control
- Rich manageability features
  - Interfaces: NC-SI, SMBUS; L2, L3 filters
  - BMC pass-thru to enable sharing the NIC with the host
- Supports Energy Efficient Ethernet (802.3 az) Adaptive Power Management



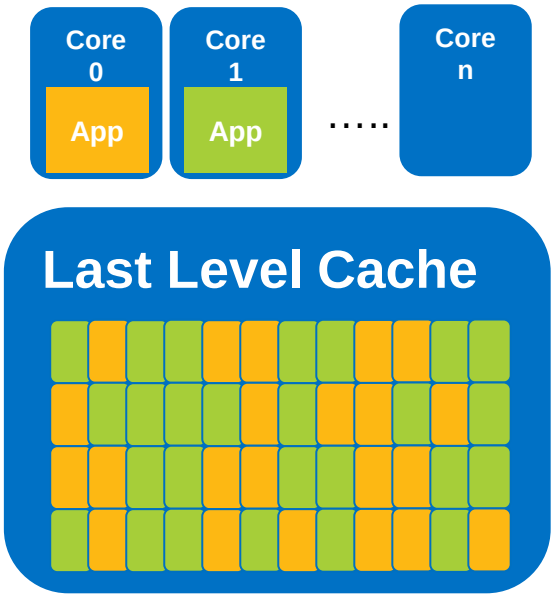
Virtualization Enhancements

- Cache Allocation Technology & Memory BW Monitoring
  - Enable the OS/VMM to monitor and manage shared platform resources on a per thread/ per VM basis
- Posted Interrupts
  - Complimentary to APIC Virtualization
  - Treats interrupts like posted mem writes; VM interrupts only when active
  - Reduces VM exits; Enables co-migration of interrupts as the VM moves
- Page Modification Logging
  - Builds on Extended Page Table A/D support on Haswell Core
  - Provides ‘dirty’ page log table to accelerate SW
- Broadwell reduces VM Entry/ Exit latency by ~20%

Feature Overview: CMT and CAT

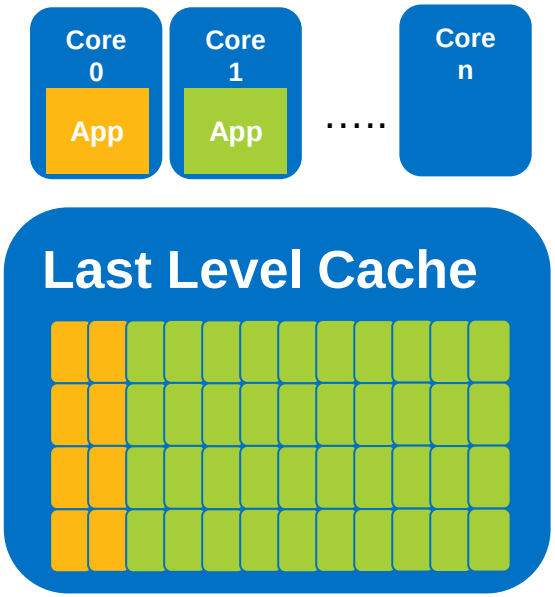
Cache Monitoring Technology (CMT)

- Identify misbehaving or cache-starved applications and reschedule according to priority
- Cache Occupancy reported on per Resource Monitoring ID (RMID) basis



Cache Allocation Technology (CAT)

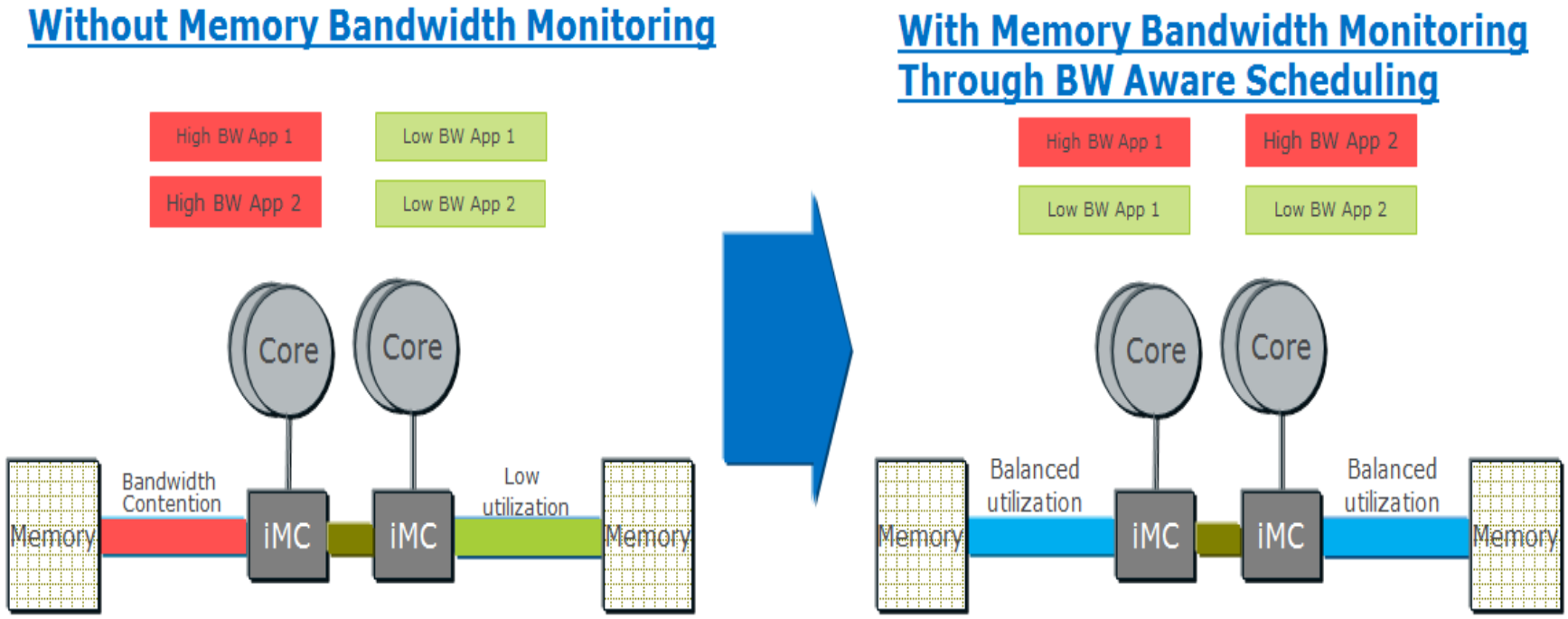
- Last Level Cache partitioning mechanism enabling the separation of applications, threads, VMs, etc.
- Misbehaving threads can be isolated to increase determinism



Cache Monitoring and Allocation Improve Visibility and Runtime Determinism

Memory Bandwidth Monitoring

Allows an OS, Hypervisor / VMM or similar system service management agent to make scheduling decisions based on memory bandwidth usage per core/thread.



Benefits/Usages

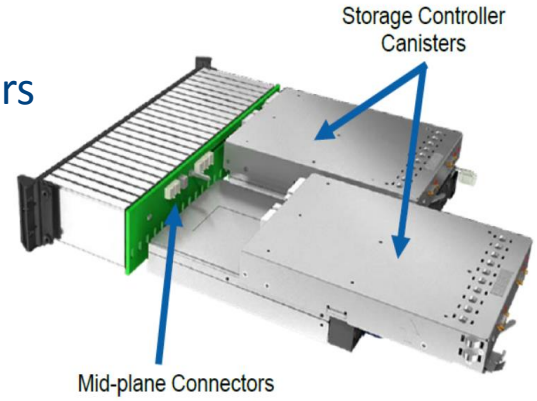
- BW-aware scheduling - Balance the BW utilization across sockets
- VM migration - Move affected or affecting VM to a different platform
- Partitioning - Feedback to cache allocation (now) and memory allocation (future) for VMs

Hardware Controlled Power Management

- Feature allows the hardware to make Power Management Decisions Autonomously
  - P and C state policies added to existing hardware mechanism
  - Utilization based algorithms used to control power-state
- Mechanism frees the OS from making frequency decisions
  - Breaks eco-system support dependency to enable feature improvements
- Hardware can make faster and more optimized decisions
  - Updates evaluated at ~1msec intervals
  - Uses granular statistical information not usable by software

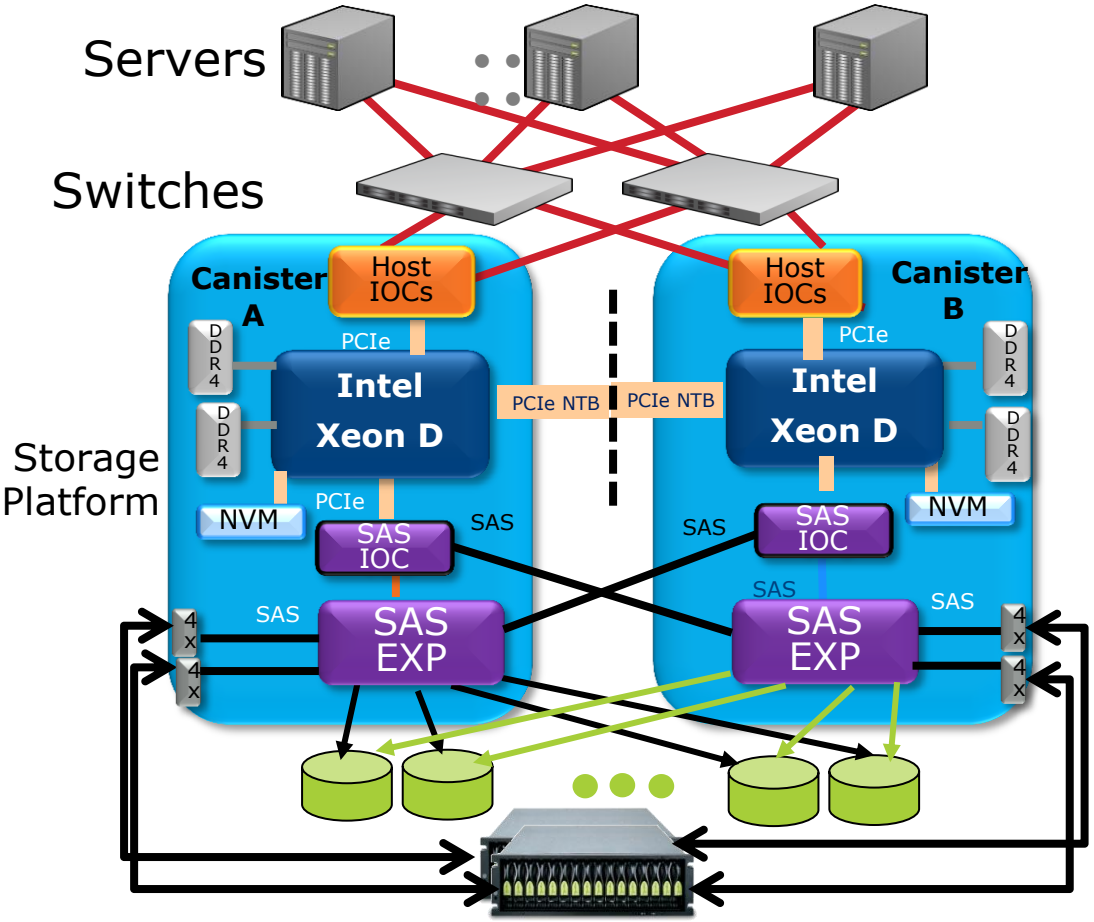
Platform Storage Extensions

- ADR – Asynchronous DRAM Refresh
  - Preserves key data in battery backed DRAM in the event of a power failure
- NTB – Non Transparent PCIe Bridging
  - 4/8/16 lanes can be configured as NTB
  - Defines “window” to remote agent memory and allows redundancy through PCIe
- Quick Data Technology
  - Provides low-latency and high throughput data transfers
  - Mem <-> Mem, Mem <-> MMIO, MMIO <-> MMIO transfers
  - Supports T10-DIF Insert/Strip/Update/Multicast
- PCIe Dual Cast
  - Allow single write transaction to multiple targets
  - Alleviates memory BW utilization for storage workloads



Storage Bridge Bay
Form Factor

Enterprise Storage Example: Dual Canister Flow

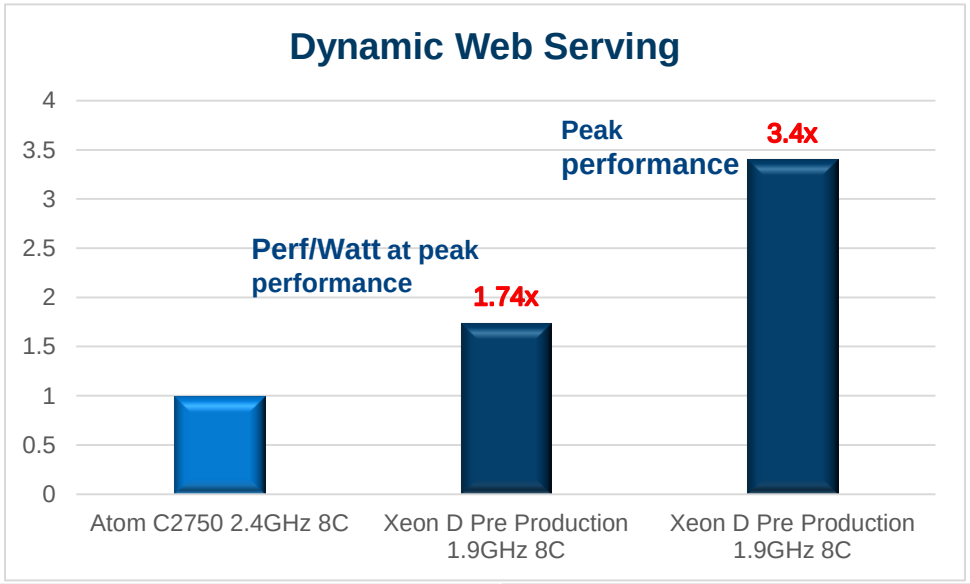


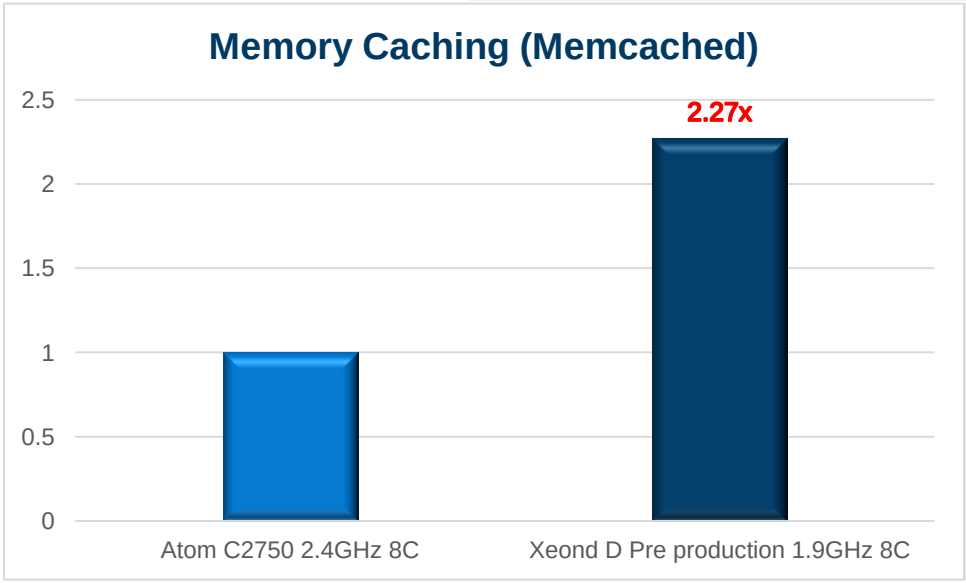
SSD: High-performance storage
SAS: Enterprise performance storage;
SATA: Enterprise bulk storage

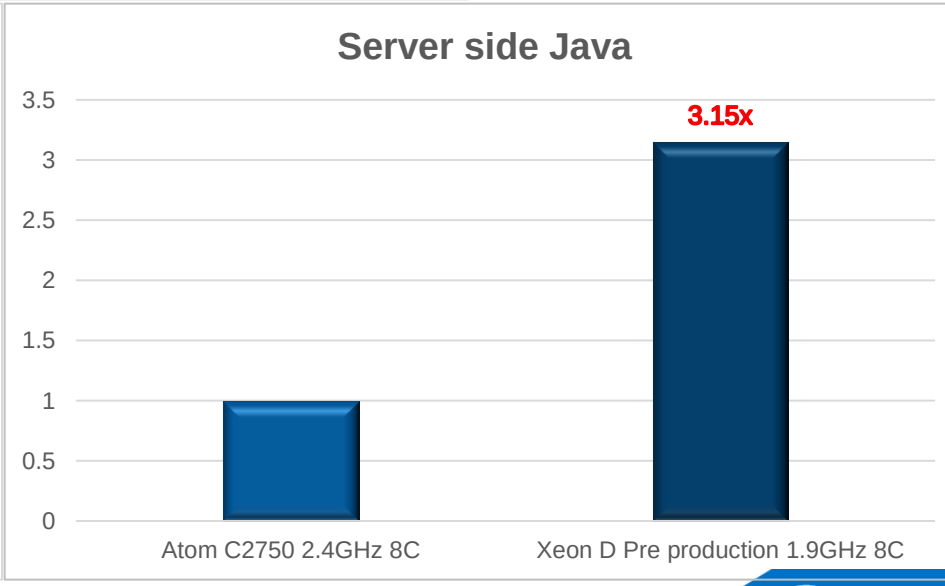
Scalable Unified Storage Solution

- Entry solutions are TDP constrained
  - Xeon D TDP range (20 -45w) critical
- Integrated Storage extensions fundamental to the solution
  - PCIe NTB, DMA engine enable higher availability
  - ADR feature for RAID Cache Data Protection
  - PCIe Dualcast reduces memory BW demand
  - Integrated Ethernet used as the host interface or for clustering
- ISA for data protection, storage efficiency and management
  - RAID-5/6, CRC, encryption, hashing and compression

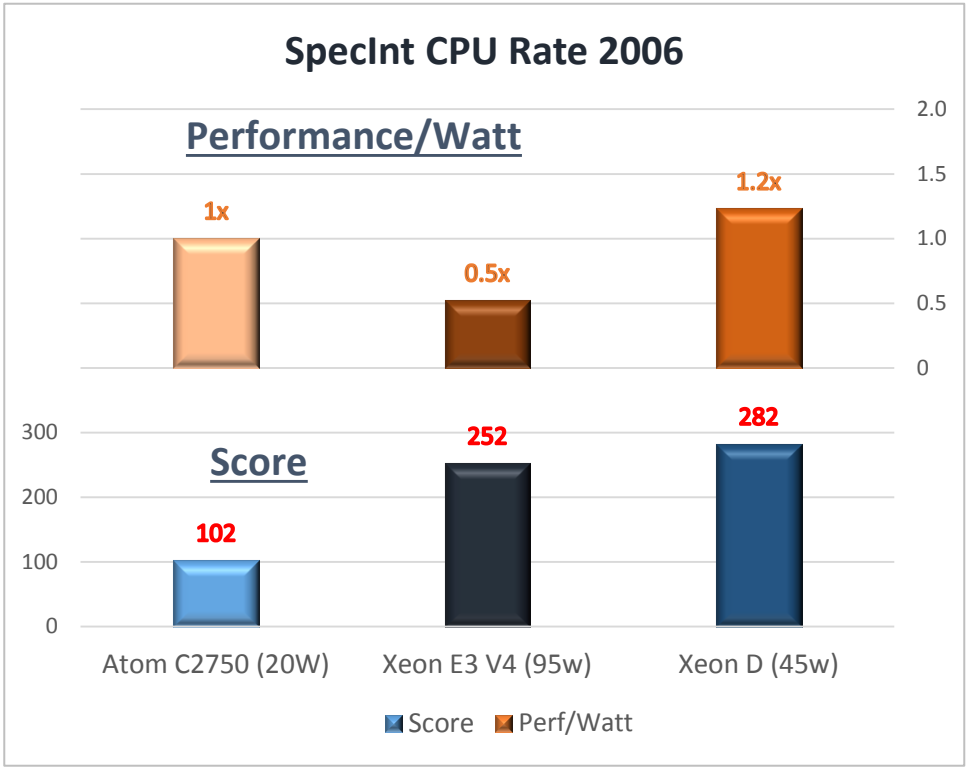
Web and Caching Tier Workloads

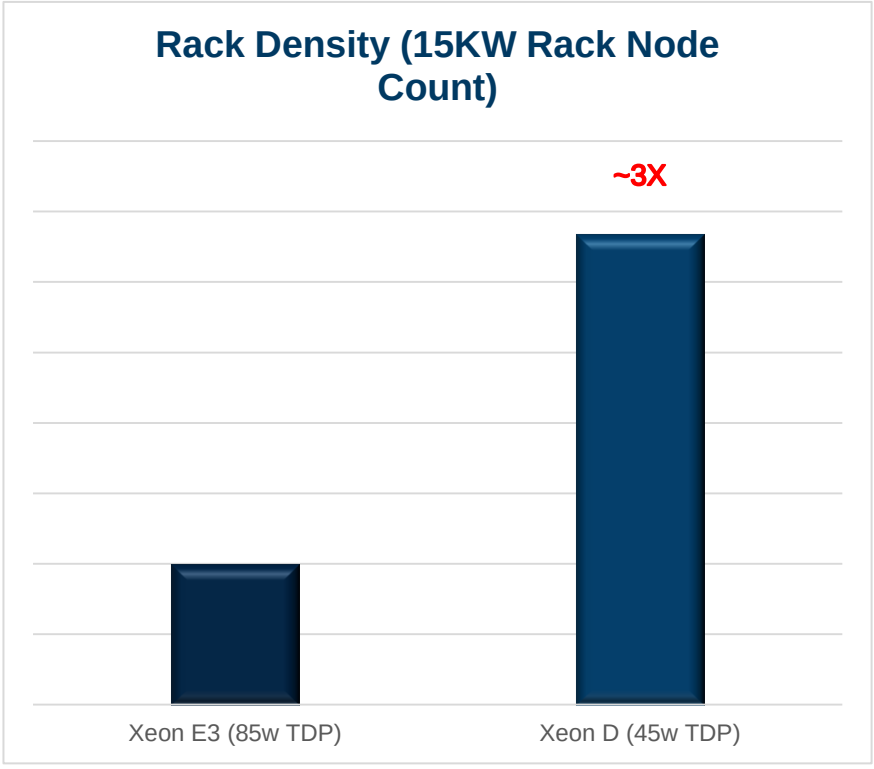






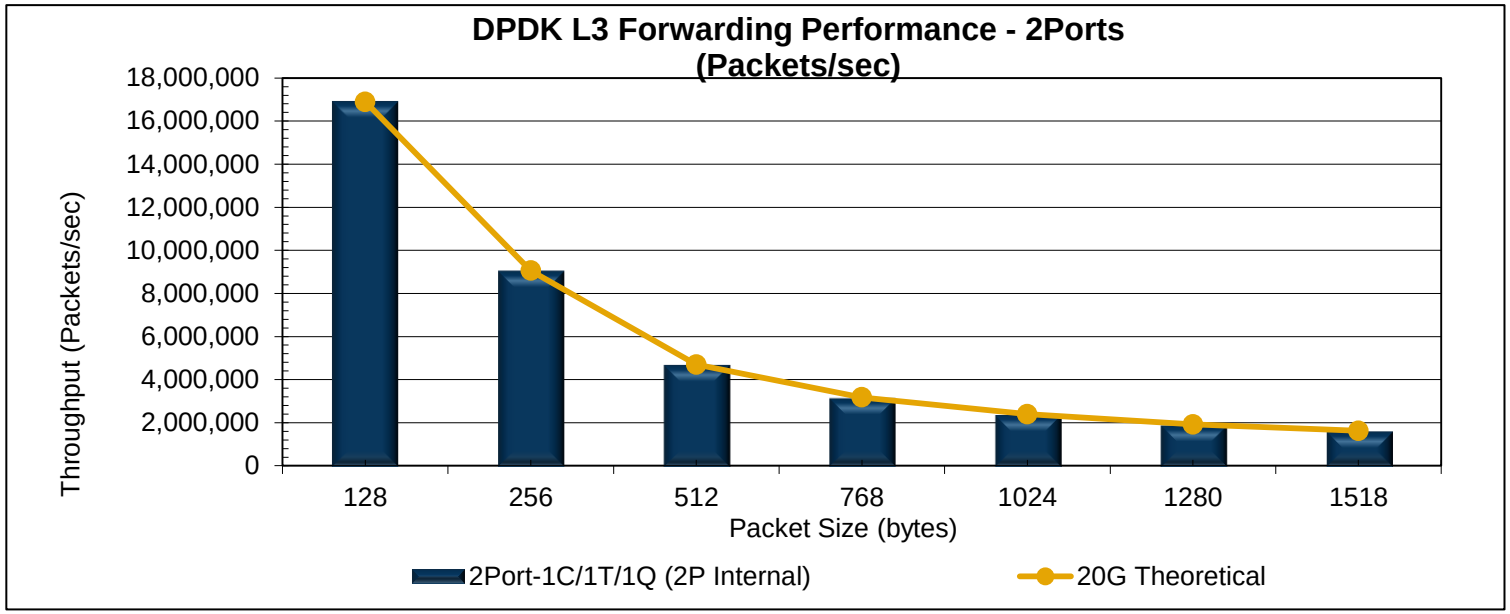
SpecInt CPU and Rack Density

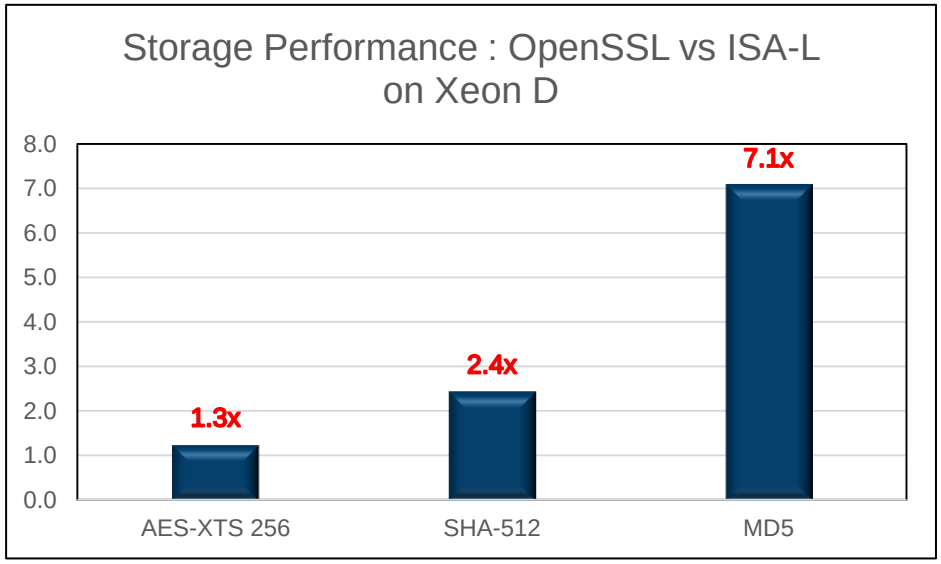




Xeon class performance at a very high power efficiency and density

Networking and Storage performance





Conclusion

- Broadwell-DE brings Xeon class performance and capabilities to dense solutions with higher power efficiency
- Focused engineering to co-optimize the platform and SoC to achieve density and power targets
- Rich feature set across Virtualization, Security RAS and Power Management
- New optimization choices for Hyperscale cloud environments, Networking and for dense, low power storage solutions
- Delivers up to 3.4X the performance and up to 1.7X perf/watt over the 22nm Atom C2000 SoC family

Glossary

AES-NI: Advanced Encryption Standard – New Instructions
BMC: Baseboard Management Controller
CA: Caching Agent
CAP: Command Address Parity
GPIO: General Purpose IO
LPC: Low Pin Count
MC: Memory Controller
MMIO: Memory Mapped IO
NC-SI: Network Controller Sideband Interface
PA: Physical Address
PCH: Peripheral Components Hub
PSE: Platform Storage Extensions
RAPL: Running Average Power Limiting
SDDC: Single DRAM Device Correct
SMBUS: System Management BUS
Socket: CPU die
SPI: Serial Peripheral Interface
Uncore: Logic on the CPU die excluding the code. Includes LLC, System Interface logic
WoL: Wake On Lan